A system and method for controlling command queuing on parity drives in an array of disk drives.

Patent Number: EP0661635, B1

Publication date: 1995-07-05

Inventor(s): ISLAM SHAH MOHAMMAD REZAUL (US)

Applicant(s):: IBM (US)

Requested Patent: JP7210334

Application Number: EP19940309362 19941214

Priority Number(s): US19930175710 19931230

IPC Classification: G06F11/10 ; G06F12/08

EC Classification: G06F11/10M, G06F12/08B12

Equivalents: DE69412775D, DE69412775T, KR162124, US5530948

Abstract

The present invention provides a technique for queuing commands on an input/output controller for a parity drive in a level 4 or level 5 redundant array of inexpensive disk drives, which responds to receipt of a write instruction with appended data by determining a logical block address of a data drive for the appended data and a logical block address of a parity drive for redundant data belonging to a destination stripe of the appended data. If a parity cache entry corresponds to the parity address, a cache hit is deemed to have occurred. Responsive to a cache hit, computation of replacement parity data for the stripe may be done from parity data in cache at a location given by the parity cache entry, data read from the data drive, and the appended data. The newly calculated parity data is then placed in cache. A command is then sent to an input/output controller for the drive where the parity data is located to write the replacement parity data. Serialization of access to the parity data is moved from the disk drive to cache memory, allowing command queuing to be used with the disk drive to preserve optimum available performance of the drive.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平7－210334

(43)公開日　平成7年（1995）8月11日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ	3/06	５４０			
		３０５　Ｃ			
	11/10	３２０　Ｅ			
	12/16	３２０　Ｌ	9293－5Ｂ		

審査請求　有　請求項の数29　ＯＬ（全 14 頁）

(21)出願番号　特願平6－236301

(22)出願日　平成6年（1994）9月30日

(31)優先権主張番号　１７５７１０

(32)優先日　1993年12月30日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

(72)発明者　シャ・モハメド・レザウル・イスラム
アメリカ合衆国33431、フロリダ州ボカ・ラトン、アパートメント　ジェイ226、ジャイウッド・テラス　3380

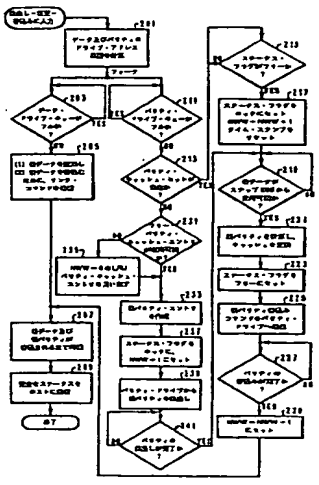
(74)代理人　弁理士　合田　潔　（外２名）

(54)【発明の名称】データ記憶方法及びキューイング方法

(57)【要約】

【目的】　集合的に１個以上の論理大容量記憶装置として機能する大容量記憶装置のセットを提供する。

【構成】　ＲＡＩＤレベル４または５のパリティ・ドライブの入出力制御装置上におけるコマンド・キューイングが、付加データを伴う書込み命令の受信に応答して、データ・ドライブの論理ブロック・アドレスと、パリティ・ドライブの論理ブロック・アドレスを決定する。パリティ・キャッシュ・エントリがパリティ・アドレスに一致すると、キャッシュ・ヒットが発生する。キャッシュ・ヒットに応答して置換パリティ・データが、パリティ・キャッシュ・エントリにより提供されるキャッシュ位置に存在するパリティ・データ、データ・ドライブから読出されるデータ、及び付加データから計算される。新たに計算されたパリティ・データがキャッシュに配置される。次に置換パリティ・データを書込むために、パリティ・データが配置されるドライブの入出力制御装置にコマンドが送信される。



1

【特許請求の範囲】

【請求項１】耐障害論理大容量記憶装置内の物理記憶装置にユーザ・データ及び冗長データを記憶する方法であって、

ホストからの付加データを伴う書込みコマンドの受信に応答して、第１の物理記憶装置上における付加データの論理ブロック・アドレスと、第２の物理記憶装置上における冗長データのパリティ・アドレスとを決定するステップと、

前記第１の記憶装置に対して、論理ブロック・アドレスの旧データを読出し、論理ブロック・アドレスに付加データを書込むように指令するステップと、

前記第２の記憶装置上の冗長データがメモリ・バッファ内に存在するかどうかを判断するステップと、

冗長データが前記メモリ・バッファ内に存在しない場合、前記第２の記憶装置から前記メモリ・バッファに冗長データを読出すステップと、

冗長データ、付加データ及び旧データから新たな冗長データを計算するステップと、

前記第２の物理記憶装置に新たな冗長データを論理ブロック・アドレスに記憶するように指令するステップと、

を含む記憶方法。

【請求項２】新たな冗長データを計算するステップが、前記メモリ・バッファ内の冗長データを新たな冗長データにより置換するステップを含む、請求項１記載の記憶方法。

【請求項３】ディスク・ドライブの冗長アレイ内のパリティ・ドライブ上にコマンドをキューイングする方法であって、

ホストからの付加データを伴う書込みコマンドの受信に応答して、データ・ドライブ上における付加データの論理ブロック・アドレスと、前記パリティ・ドライブ上におけるパリティ・データの論理ブロック・アドレスとを決定するステップと、

パリティ・キャッシュ・エントリがパリティ・データの論理ブロック・アドレスに一致するかどうかを判断するステップと、

前記一致の肯定判定に応答して、パリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションのパリティ・データと、前記データ・ドライブから読出されるデータと付加データとから、置換パリティ・データを計算するステップと、

置換パリティ・データを前記バッファ・ロケーションに配置するステップと、

置換パリティ・データを書込むためのコマンドを、パリティ・データの論理ブロック・アドレスに対応する入出力制御装置に送信するステップと、

を含むキューイング方法。

【請求項４】前記一致の否定判定に応答して、パリティ・データを前記バッファ・ロケーションに読出すためのコマンドを、前記パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信するステップと、

前記パリティ・ドライブの論理ブロック・アドレスのパリティ・キャッシュ・エントリを生成するステップと、

を含む、請求項３記載のキューイング方法。

【請求項５】パリティ・キャッシュ・エントリが前記パリティ・ドライブの論理ブロック・アドレスと、ドライブ識別と、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、メモリ・バッファ内のパリティ・データの位置を示す位置フィールドとを含む、請求項４記載のキューイング方法。

【請求項６】前記データ・ドライブの論理ブロック・アドレスの決定に応答して、データを読出し次に付加データを書込むためのリンク・コマンドを、論理ブロック・アドレスをカバーする入出力制御装置に送信するステップ、

を含む、請求項５記載のキューイング方法。

【請求項７】計算ステップがリンク・コマンドの送信ステップの読出しオペレーションの後に発生する、請求項６記載のキューイング方法。

【請求項８】前記パリティ・キャッシュ・エントリの生成ステップが、

パリティ・キャッシュ・テーブルが未使用のパリティ・キャッシュ・エントリを有するかどうかを判断するステップと、

肯定判定に応答して、未使用のパリティ・キャッシュ・エントリを使用するステップと、

否定判定に応答して、待機中の書込みを有さないパリティ・キャッシュ・エントリの中から、最低使用頻度のパリティ・キャッシュ・エントリを探し出し、探し出されたパリティ・キャッシュ・エントリを使用するステップと、

を含む、請求項５記載のキューイング方法。

【請求項９】置換パリティ・データ書込みのためのコマンドの送信ステップが、書込み待機カウンタを増分するステップを含む、請求項５記載のキューイング方法。

【請求項１０】置換パリティ・データの生成のために、前記バッファ・ロケーション内のパリティ・データへのアクセスを直列化するステップ、

を含む、請求項９記載のキューイング方法。

【請求項１１】置換パリティ・データが前記パリティ・ドライブの論理ブロック・アドレスに書込まれた後、パリティ・キャッシュ内の書込み待機カウンタを減分するステップ、

を含む、請求項１０記載のキューイング方法。

【請求項１２】前記パリティ・ドライブがチェック・ディスクである、請求項１１記載のキューイング方法。

【請求項１３】前記パリティ・ドライブがディスク・ド

ライブの冗長アレイの任意の１つに相当する、請求項１１記載のキューイング方法。

【請求項１４】ホスト・データ処理システムからの付加データを伴う書込み命令の受信に応答して、データ・ドライブ上における付加データの論理ブロック・アドレスと、パリティ・ドライブ上における冗長データの論理ブロック・アドレスとを決定する手段と、

複数のパリティ・キャッシュ・エントリを有するパリティ・キャッシュ・テーブルと、パリティ・キャッシュ・エントリ及び前記パリティ・ドライブ上の論理ブロック・アドレスに対応するパリティ・データを記憶するバッファと、

パリティ・キャッシュ・エントリが発見手段により見い出される前記パリティ・ドライブの論理ブロック・アドレスに一致するかどうかを判断する手段と、

前記一致の肯定判定に応答して、パリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションのパリティ・データと、前記データ・ドライブの論理ブロック・アドレスから読出されるデータと、付加データとから、置換パリティ・データを計算する手段と、

置換パリティ・データを前記バッファ・ロケーションに配置する手段と、

置換パリティ・データを書込むためのコマンドを、パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段と、

を含む、ディスク・ドライブの冗長アレイ。

【請求項１５】前記一致の否定判定に応答して、パリティ・データを前記バッファ・ロケーションに読出すためのコマンドを前記パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段と、

前記パリティ・ドライブの論理ブロック・アドレスから読出されるパリティ・データに対応するパリティ・キャッシュ・エントリを生成する手段と、

を含む、請求項１４記載のディスク・ドライブの冗長アレイ。

【請求項１６】パリティ・キャッシュ・エントリが前記パリティ・データの論理ブロック・アドレスと、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、ドライブ識別と、メモリ・バッファ内のパリティ・データの位置フィールドとを含む、請求項１５記載のディスク・ドライブの冗長アレイ。

【請求項１７】前記データ・ドライブの論理ブロック・アドレスの決定に応答して、データを読出し次に付加データを書込むためのリンク・コマンドを、前記データ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段、

を含む、請求項１６記載のディスク・ドライブの冗長アレイ。

【請求項１８】パリティ・キャッシュ・エントリの生成

手段が、

パリティ・キャッシュ・テーブルが未使用のパリティ・キャッシュ・エントリを有するかどうかを判断する手段と、

肯定判定に応答して、未使用のパリティ・キャッシュ・エントリを使用する手段と、

否定判定に応答して、待機中の書込みを有さない最低使用頻度のパリティ・キャッシュ・エントリを探し出し、探し出されたパリティ・キャッシュ・エントリを使用する手段と、

を含む、請求項１６記載のディスク・ドライブの冗長アレイ。

【請求項１９】置換パリティ・データ書込みのためのコマンドの送信手段が、書込み待機カウンタを増分する手段を含む、請求項１６記載のディスク・ドライブの冗長アレイ。

【請求項２０】置換パリティ・データの生成のために、前記バッファ・ロケーション内のパリティ・データへのアクセスを直列化する手段、

を含む、請求項１９記載のディスク・ドライブの冗長アレイ。

【請求項２１】置換パリティ・データが前記パリティ・ドライブの論理ブロック・アドレスに書込まれた後、前記パリティ・ドライブの論理ブロック・アドレスに対応するパリティ・キャッシュ内の書込み待機カウンタを減分する手段、

を含む、請求項２０記載のディスク・ドライブの冗長アレイ。

【請求項２２】前記パリティ・ドライブがチェック・ディスクである、請求項２１記載のディスク・ドライブの冗長アレイ。

【請求項２３】前記パリティ・ドライブがディスク・ドライブの冗長アレイの任意の１つに相当する、請求項２１記載のディスク・ドライブの冗長アレイ。

【請求項２４】論理装置として動作する物理記憶装置のアレイと、

物理記憶装置のアレイ上の所定のロケーションに記憶される付加データを伴う更新コマンドを提供するコマンド発生器と、

ユーザ・データのストライプに対応する選択パリティ・データ・ブロックのコピーを記憶するキャッシュと、

ユーザ・データ・ブロック及びユーザ・データ・ブロックのストライプに渡って生成されるパリティ・データ・ブロックを物理記憶装置のアレイ間でストライプし、物理記憶装置ロケーションによりキャッシュ内に記憶される選択パリティ・データ・ブロックの１個以上のコピーを識別するパリティ・キャッシュ・テーブルを保守するローカル・プロセッサと、

ローカル・プロセッサによりそれぞれの入出力制御装置に送信されるコマンドをキューイングする、物理記憶装

5

置の物理アレイの各々に関連する入出力制御装置と、
を含むデータ処理システム。
【請求項２５】パリティ・キャッシュ・エントリが物理記憶装置の論理ブロック・アドレスと、物理記憶装置の識別と、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、キャッシュ内のパリティ・データ・ブロックの位置フィールドとを含む、請求項２４記載のデータ処理システム。
【請求項２６】ローカル・プロセッサが、キャッシュ内の選択パリティ・データ・ブロックの各コピーへのアクセスを制御するシリアライザ、
を含む、請求項２５記載のデータ処理システム。
【請求項２７】キャッシュ内にパリティ・データ・ブロックが存在しないストライプ内の物理アドレスを更新するコマンドに応答して、前記ストライプに対応するパリティ・データ・ブロックをキャッシュに読出し、パリティ・キャッシュ・テーブルに前記パリティ・データ・ブロックのエントリを生成するパリティ・キャッシュ・テーブル・エントリ発生器、
を含む、請求項２６記載のデータ処理システム。
【請求項２８】パリティ・データが単一の物理記憶装置上に配置される、請求項２７記載のデータ処理システム。
【請求項２９】パリティ・データが全ての物理記憶装置上に配置される、請求項２７記載のデータ処理システム。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は集合的に１個以上の論理大容量記憶装置として機能する大容量記憶装置のセットに関し、特に、ＲＡＩＤレベル４及び５システムにおけるパリティ・ドライブ上で、コマンドをキューに待機させる（以降ではコマンド・キューイングと記す）システム及び方法に関する。
【０００２】
【従来の技術】ディスク・メモリの使用はコンピュータにおいて重要であり続ける。なぜなら、これは不揮発性であり、メモリ・サイズ要求が主メモリの実際の容量を上回り続けるからである。ディスクは主メモリよりも低速であるので、システム性能はしばしばディスク・アクセス速度により制限される。従って、全体的システム性能にとって、メモリ・サイズ及びディスク・ドライブ・ユニットのデータ・アクセス速度を増大することが重要である。この議論に関しては、Michelle Y. Kimによる"Synchronized Disk Interleaving" (IEEE Transactions On Computers、Vol. C-35、No. 11、１９８６年１１月）を参照されたい。
【０００３】ディスク・メモリ・サイズはディスクの数を増やしたり、ディスクの直径を大きくすることにより

6

増大するが、データ・アクセス速度は向上しない。メモリ・サイズ及びデータ転送レートの両方が、データ記憶密度を増加することにより増大される。データ転送レートはディスクの回転速度を増すことにより向上する。しかしながら、技術的制約がデータ密度を制限し、高密度且つ高速ディスクはエラーを生じ易い。
【０００４】データ・アクセス速度を改良するために、様々な技術が使用されてきた。単一トラック上のデータへの連続アクセスにおけるシーク及び回転遅延を排除するために、データの全トラックを保持可能なディスク・キャッシュ・メモリが使用されてきた。ディスク・セット上のまたは単一ディスクのトラック・セット上のデータ・ブロックをインタリーブするために、複数の読出し／書込みヘッドが使用されてきた。一般的データ・ブロック・サイズはバイト・サイズ、ワード・サイズ及びセクタ・サイズである。ディスク・インタリービングは性能を向上させるための既知のスーパーコンピュータ技術であり、前記記事などにおいて述べられている。
【０００５】データ・アクセス性能は関連出願に述べられるように、多数のパラメータにより測定される。トランザクション処理（バンキングなど）では、データ転送は、通常、小さく、要求レートは高速且つランダムである。一方、スーパーコンピュータの応用では、大きなデータ・ブロックの転送が一般的である。
【０００６】最近開発された比較的低コストで性能が改良されたインタリーブ式ディスク・メモリ・アーキテクチャは、ＲＡＩＤ (Redundant Arrays of Inexpensive Disk) としてグループ化される。例えばDavid A. Pattersonらによる"A Case forRedundant Arrays of Inexpensive Disks（ＲＡＩＤ）" (Report No. UCB/CSD 87/89、December、1987、Computer Science Division（ＥＥＣＳ）、Universityof California、Berkeley、California 94720) を参照されたい。Pattersonらによる参考文献において述べられるように、大きなパーソナル・コンピュータ市場が、単一大型高価ディスク（ＳＬＥＤ：Single Large Expensive Disk ）システムに勝る性能対コスト比を有する安価な（inexpensive ）ディスク・ドライブの開発を支えてきた。安価なディスクにおける１読出し／書込みヘッド当たりの１秒当たりの入出力数は、大型ディスクの２倍以内である。従って、安価ディスクのセットが単一の論理ディスク・ドライブとして機能するＲＡＩＤアーキテクチャにおける、幾つかの安価ディスクからの並列転送は、低価格において、ＳＬＥＤに勝る性能を提供する。
【０００７】しかしながら、データが複数のディスク上に記憶される時、平均故障寿命（ＭＴＴＦ：mean time to failure）はアレイ内のディスクの数の増加に相反して減少する。システムのこの減少する平均故障寿命を修正するために、誤り検出及び訂正が全てのＲＡＩＤアーキテクチャの特徴となる。Patterson らの参考文献は、

各々が誤り検出及び訂正のための異なる手段を有する５つのＲＡＩＤアーキテクチャについて述べている。これらのＲＡＩＤアーキテクチャは、ＲＡＩＤレベル１乃至５として参照される。

【０００８】ＲＡＩＤレベル１はデータの完全な複製（"ミラーリング（mirroring）"とも呼ばれる）を使用し、１ディスク当たり比較的低い性能比率を有する。ＲＡＩＤレベル２は、エラー訂正及びディスク故障回復を提供するために必要な余分なディスク数を低減するエラー訂正コードを使用することにより、１ディスク当たりの容量比率に加え、この性能を改良する。ＲＡＩＤレベル２では、データはＧ個のデータ・ディスクのグループにインタリーブされ、単一のエラーを検出し訂正するために、エラー訂正コード（ＥＣＣ）が生成され、これが"チェック・ディスク"として参照されるＣ個の追加のディスクのセットに記憶される。ＥＣＣはデータ内のランダムな単一ビット・エラーを検出し訂正するために使用され、またＧ個のデータ・ディスクの１個がクラッシュする場合のデータの回復を可能とする。Ｃ＋Ｇ個のディスクの内のＧ個だけがユーザ・データを転送するので、１ディスク当たりの性能はＧ／（Ｇ＋Ｃ）に比例する。Ｇ／Ｃは、通常、１よりも相当に大きく、従ってＲＡＩＤレベル２はＲＡＩＤレベル１よりも１ディスク当たりの性能を改良する。１個以上のスペア・ディスクがシステム内に含まれ、ディスク・ドライブの１個が故障すると、このスペア・ディスクが電子的にＲＡＩＤ内にスイッチされ、故障したディスク・ドライブを置換することも可能である。

【０００９】ＲＡＩＤレベル３はＲＡＩＤレベル２の変形であり、チェック・ディスク数を１個に減らすために、ほとんどの既存の安価ディスク・ドライブにより提供されるエラー検出機能が使用され、それによりＲＡＩＤレベル２に比較して、１ディスク当たりの相対性能を向上させる。通常、パリティ・データがＥＣＣの代わりに代用される。ＥＣＣまたは他のエラー・コード、或いはパリティ・データは冗長データと称される。ＲＡＩＤレベル２及び３の両方において、大きなデータまたはグループ化データに対するディスク・アクセスのトランザクション時間は低減される。なぜなら、全てのデータ・ディスクに対する帯域幅が利用されるからである。

【００１０】例えばトランザクション処理において一般的である小さなデータ転送の性能基準は、ＲＡＩＤレベル１乃至３に対して不十分であることが知られている。なぜなら、データがディスク間でビット・サイズまたはバイト・サイズのブロックにインタリーブされ、１セクタより少ないデータへのアクセスにおいてさえも、全てのディスクがアクセスされなければならないからである。この性能パラメータを改良するために、ＲＡＩＤレベル３の変形であるＲＡＩＤレベル４では、レベル１乃至３におけるビットまたはバイト・インタリーブ・モードの代わりに、データがセクタ・インタリーブ・モードによりディスクにインタリーブされる。換言すると、個々の入出力転送が単一のデータ・ディスクだけに関連する。これによる利点は、入出力オペレーションの並列性の潜在に由来する。これは同一のデータ・ディスクを同時にアクセスする別々のデータ・アクセス要求の間の競合を低減する。

【００１１】それにも関わらず、ＲＡＩＤレベル４の性能は、書込みオペレーションの間のチェック・ディスクに対するアクセス競合のために制限される。全ての書込みオペレーションにおいて、データが書込まれる各ストライプ（すなわちセクタの行）に対応して、チェック・ディスク上に更新パリティ・データを記憶するために、チェック・ディスクがアクセスされなければならない。Patterson らは、ＲＡＩＤレベル４及び５において、単一のセクタに対する個々の書込みが、論理大容量記憶装置内の全てのディスクに関連しないことを観測した。なぜなら、チェック・ディスク上のパリティ・ビットが単にグループ内の全ての対応データ・ビットの単一の排他的論理和に相当するからである。ＲＡＩＤレベル４では、書込みオペレーションは常にパリティ・ディスクの読出し及び再書込みを含み、パリティ・ディスクが低カレント書込みオペレーションにおけるアレイへのアクセスのボトルネックとなる。ＲＡＩＤレベル４の変形であるＲＡＩＤレベル５は、パリティ・チェック・データとユーザ・データを全てのディスクに渡って分散することにより、書込みオペレーションの競合問題を軽減する。ＲＡＩＤレベル４では、大きな書込みオペレーション（全てのパリティ・ストライプ・ユニットに広がる）は予備読出しを要求しない。

【００１２】しかしながら、競合問題は依然として発生する。ＲＡＩＤレベル４及び５は共に、各読出し－変更－書込みオペレーション（例えばレコードの更新）において、２個の各ディスクに２回のアクセスを要求する。更新はデータ・ディスク上の既存のユーザ・データの読出しと、ユーザ・データが帰属するストライプに対応するパリティ・ディスク上のパリティ・データの読出しとを含む。これに続き、更新されたユーザ・データ及びパリティ・データが、それぞれ両方のディスクに書込まれる。読出しオペレーションは更新パリティを計算するのに不可欠であり、次の関数を用いて実行される。

新パリティ＝（旧データ．ＸＯＲ．新データ）．ＸＯＲ．旧パリティ

【００１３】パリティ・データのコヒーレンシの喪失を阻止するために、ＲＡＩＤレベル４及び５の大容量記憶システムにおけるデータ更新オペレーション処理は、原子的または直列化読出し変更書込みオペレーションを要求し、その間にパリティ・データを記憶するドライブがロックされ、最初の更新が完了する以前に、別のデータ更新オペレーションのパリティ情報が変化するのを阻止

する。パリティのコヒーレンシとは、パリティ・グループのデータに対して順次的に実行される一連の排他的論理和演算を、パリティが継続的に表すことを意味する。ドライブのロックは、タグド・コマンド・キューイング（ＴＣＱ：Tagged Command Queuing）をサポートするディスク・サブシステムにおいて、コマンドのキューイングを阻止する。

【００１４】タグド・コマンド・キューイングはＳＣＳＩ（Small Computer SystemsInterface）の規格において定義される。これは応答を待機することなく、ホストによりドライブに送信される複数のコマンドを処理する。コマンド及び応答は、ホストが応答を要求にマッチさせるようにタグ付けされる。幾つかのシステムでは、ドライブ性能を改良するためにオペレーションの実行順序の最適化が行われる。必要に応じて、所定の順序によるコマンドの実行を保証するために、リンク化コマンドが提供される。ドライブに対するアクセスの直列化は、コマンド・キューイングを阻止するので、その後、ディスク・サブシステム制御装置はオペレーション・シーケンスを最適化できず、ディスク・サブシステムの性能に大きく影響を及ぼす結果となる。

【００１５】用語"ストライピング（striping）"がＲＡＩＤ技術においてしばしば見受けられる。ストライピングは"ストライプ・ユニット"により、複数のディスク・ドライブに渡ってデータをインタリーブすることである。ストライプ・ユニットは論理的に連続なデータのグループであり、これはデータを異なるディスク上に配置する以前に、単一のディスク上に物理的に連続に書込まれる。

【００１６】

【発明が解決しようとする課題】本発明の目的は、集合的に１個以上の論理大容量記憶装置として機能する大容量記憶装置のセットを提供することである。

【００１７】本発明の別の目的は、ＲＡＩＤレベル４及び５システムにおけるパリティ・ドライブ上で、コマンドをキューイングするシステム及び方法を提供することである。

【００１８】

【課題を解決するための手段】本発明の上述及び他の目的が次のようにして達成される。ディスク・ドライブのレベル４またはレベル５冗長アレイ内のパリティ・ドライブの入出力制御装置にコマンドをキューイングする方法が、付加データを伴う書込み命令の受信に応答して、付加データに対応するディスク・ドライブ上の論理ブロック・アドレスと、付加データの宛先ストライプに帰属する冗長データに対応するディスク・ドライブ上の論理ブロック・アドレスを決定する。パリティ・キャッシュ・エントリがパリティ・データ論理ブロック・アドレスに一致すると、キャッシュ・ヒットが発生する。キャッシュ・ヒットに応答して、ストライプに対応する置換パリティ・データがパリティ・キャッシュ・エントリにより提供されるキャッシュ・ロケーションに存在するパリティ・データ、データ・ドライブから読出されるデータ及び付加データから計算される。新たに計算されたパリティ・データが次にキャッシュに配置される。次に置換パリティ・データを書込むために、パリティ・データが配置されるドライブの入出力制御装置にコマンドが送信される。パリティ・データに対するアクセスの直列化が、ディスク・ドライブからキャッシュ・メモリに転送され、それによりパリティ・データに対応するディスク・ドライブにおいてコマンド・キューイングが使用可能となり、高性能オペレーションが確保される。

【００１９】

【実施例】図１を参照すると、データ処理システム１１が示される。データ処理システム１１はシステム中央処理ユニット（ＣＰＵ）１３、システム・メモリ１５、大容量記憶制御装置１７、及び通信リンク１９を含み、通信リンク１９はＣＰＵ１３、システム・メモリ１５及び大容量記憶制御装置１７をリンクし、それらの間のデータ及びコマンドの交換を可能とする。通信リンク１９はシステム・バスまたはあるタイプのネットワークを表す。

【００２０】大容量記憶制御装置１７はデータ・ブロックをＲＡＩＤ２１にストライプし、またＲＡＩＤ２１からデータ・ブロックを回復する機能を提供する。本発明はレベル４またはレベル５モードのＲＡＩＤ２１の使用に適応する。大容量記憶制御装置１７は通信リンク１９とのインタフェース２３を含む。インタフェース２３はシステム・バス（例えばマイクロチャネル、ＥＩＳＡなど）またはＳＣＳＩ接続、或いは通信リンク１９へのネットワーク・アダプタである。ローカル・プロセッサ２５とメモリ・バッファ２７との間には、排他的論理和プロセッサ２９が接続される。排他的論理和プロセッサ２９は、ＲＡＩＤ２１において冗長情報として使用されるパリティ・データをデータ・ストライプに渡って生成し、性能を改良するために使用される。排他的論理和プロセッサ２９はまたパリティ情報の更新を提供する。

【００２１】様々なタイプのデータがメモリ・バッファ２７上に記憶され、そうしたデータにはパリティ・キャッシュ・テーブル３１及びパリティ・データ・ブロック３３が含まれる。ユーザ・データ・ブロック３５は、特にエラー回復のために遷移してメモリ・バッファ２７を通過する。ローカル・プロセッサ２５は通信リンク１９からインタフェース２３を介して受信されるデータを取得し、ＲＡＩＤ２１内の複数の直接アクセス記憶装置間でデータをストライプするように編成する。ローカル・プロセッサ２５はまた、メモリ・バッファ２７を用いてＲＡＩＤ２１からデータを回復及び再編成し、インタフェース２３を介してデータ処理システム１１の計算ユニットに提供する。

【００２２】ローカル・プロセッサ２５はまた、メモリ・バッファ２７内のパリティ・データ・ブロック３３へのアクセスを直列化する処理を実施するために使用される。データは大容量記憶制御装置１７から複数の入出力制御装置３７Ａ乃至３７Ｄを介して、ＲＡＩＤ２１に渡される。各入出力制御装置３７Ａ乃至３７Ｄはそれぞれローカル・バッファ３９Ａ乃至３９Ｄをアクセスし、直接アクセス記憶装置またはディスク・ドライブ４１Ａ乃至４１Ｄをそれぞれ制御する。各ローカル・バッファ３９Ａ乃至３９Ｄ内には、それぞれコマンド・キュー４３Ａ乃至４３Ｄが存在する。入出力制御装置３７Ａ乃至３７Ｄは、コマンド・キュー４３Ａ乃至４３Ｄ内に待機されるコマンドの実行順序を特定の制約内において最適化する。

【００２３】コンピュータ・ベースのデータ処理システム１１のシステム・メモリ１５はオペレーティング・システム４５を含み、これには大容量記憶制御装置１７により実現される論理記憶装置のためのデバイス・ドライバを含む複数のデバイス・ドライバが含まれる。システム・メモリ１５には更に、ＣＰＵ１３により大容量記憶制御装置１７に送信されるコマンドのためのデバイス・ドライバ・キューが含まれる。

【００２４】図２はディスク・ドライブ１２０の上面図であり、本発明を理解する上で有用なディスク・ドライブ・オペレーションの態様を表す。ディスク１４２の磁気表面１４０上の複数のトラック（１乃至Ｎ＋１）の１つに対するアーム１２３によるスライダ１２６の位置決めが、トラック内のセクタ（セクタ１４６及び１４８など）からデータを読出しまたは書込むために実行される。各トラックはディスク１４２の中心から放射状に広がる複数のトラック・サーボ・フィールド１４４により、セグメントまたはセクタに分割される。トラック・サーボ・フィールドは、回転アーム１２３の端部のスライダ１２６の移動に一致するように湾曲される。ディスク回転スピードが一定の場合（すなわち一定角速度または"ＣＡＶ"（constant angular velocity））、スライダ１２６に実装されるトランスジューサがトラック・サーボ・フィールド１４４に厳密な時間間隔で遭遇する。トラッキング情報が既知のように、トラック・サーボ・フィールド１４４から導出される。スライダ１２６はサーボ・フィールドの間を盲目的に浮上する。ユーザ・データまたはパリティ・データがディスク・ドライバ１２０上で更新される時、旧データが所定の位置から読出され、置換データが常にそこに書込まれる。例えばセクタ１４６がどのようなタイプのデータを含んでいようと、これは最初の回転において読出され、更新データが第２のまたは以降の回転においてセクタに書込まれる。セクタに対する読出しと書込みとの間に少なくとも１回転に対応する遅延が生じる。従来技術では、この間にパリティ・データを記憶するディスクへのアクセスが許可されず、ドライブ性能に影響を与えた。

【００２５】図３は４つのディスク・ドライブ間における、ユーザ・データ及びパリティ・データ・ブロックのストライピングを示す。ＲＡＩＤレベル４及びＲＡＩＤレベル５（左対称）の両方が表される。ＲＡＩＤレベル４では、４つのストライプが示され、第１のストライプはドライブ０乃至３上にそれぞれデータ・ブロックＤ０、Ｄ１、Ｄ２及びＰ０を含み、第２のストライプはドライブ０乃至３上にそれぞれデータ・ブロック３乃至５及びパリティ・ブロック１を含み、第３のストライプはドライブ０乃至３上にそれぞれデータ・ブロック６乃至８及びパリティ・ブロック２を含み、第４のストライプはドライブ０乃至３上にそれぞれデータ・ブロック９乃至１１及びパリティ・ブロック３を含む。ＲＡＩＤレベル４では、第３のドライブはチェック・ディスクとして知られる。

【００２６】ＲＡＩＤレベル５では、任意のドライブがパリティ・データを含むために、チェック・ディスクを含まない。ＲＡＩＤレベル５の一例では、ＲＡＩＤレベル４と全く同じメンバシップを有するストライプを有する。しかしながら、第１のストライプ以降ではデータ分布は異なる。第２のストライプでは、パリティ・ブロックがドライブ２に移行し、第３のストライプでは、パリティ・ブロックがドライブ１に移行し、第４のストライプでは、パリティ・ブロックがドライブ０に移行する。第２のストライプでは、データ・ブロックが左回転され、端のブロックがドライブ３に移動する。第３のストライプでは、ブロックの平均変位は左方向に２ドライブとなり、４番目のストライプでは左方向に３ドライブとなる。ＲＡＩＤレベル５は、チェック・ディスクへのアクセスの競合を回避するように設計される。

【００２７】図４は、ドライブ１、３及び４のコマンド・キュー及びデバイス・ドライバ・キュー上でのコマンドのキューイングを表すタイミング図である。図４で考慮される例は、ＲＡＩＤレベル４またはＲＡＩＤレベル５に適用され、ホストからの書込み更新コマンド、すなわちドライブ０上のブロックＤ０を最初に更新し、次にドライブ２上のブロックＤ２を更新するように指示するコマンドに関連する。これらの更新の各々は同一のストライプに対応するパリティの更新を要求し、これはどちらの場合にもパリティ・ブロックＰ０に相当する。従って、Ｐ０へのアクセスの競合が発生する。更新のための各ホスト要求は更新を構成する付加データを含むか、データを検索するためのシステム・メモリ１５内の位置を識別する。時刻Ｔ１において、データ・ブロック０にデータを書込むためのコマンドがデバイス・ドライバ・キュー４７に存在する。時刻Ｔ２において、ブロック０内のデータを読出し、付加データをブロック０に書込むためのリンク対コマンドがドライブ０のコマンド・キューに現れる。また時刻Ｔ２では、パリティ・ブロック０を

読出すための命令がドライブ３のキューに配置される。これはパリティ・ブロック０のデータがメモリ・バッファ２７に保持されていることを示すキャッシュ・ヒットが発生しない場合に限り、発生する。

【００２８】時刻Ｔ３において、データ・ブロックＤ２のデータを読出すコマンドがデバイス・ドライバ・キュー４７に配置される。その後時刻Ｔ４において、データ・ブロックＤ２を読出すコマンドから導出されるリンク・コマンドが、ドライブ２のコマンド・キューに現れる。これにはデータ・ブロックＤ２の読出し及びデータ・ブロックＤ２の書込みが含まれる。また時刻Ｔ４では、パリティ・ブロックＰ０を読出す命令がドライブ３のコマンド・キューを通じて進行するように示される。Ｔ４とＴ６との間のある時刻において、パリティ・ブロックＰ０がメモリ・バッファ２７に読出される。また時刻Ｔ６より先にデータ・ブロックＤ０を読出す命令が実行される。時刻Ｔ６において、更新データをデータ・ブロックＤ０に書込むコマンドが、ドライブ０のコマンド・キューの出力端に達するように示される。このコマンドは時刻Ｔ１２より以前に実行される。時刻Ｔ７において、パリティ・ブロックＰ０のデータを置換する書込み命令が、ドライブ３のコマンド・キューに配置される。このコマンドの進行は時間間隔Ｔ８を経過して継続される。時刻Ｔ９において、ユーザ・データ・ブロックＤ２の読出し命令がドライブ２のコマンド・キューから実行され、更新データをデータ・ブロックＤ２に書込む命令がコマンド・キューの出力端に達する。実行はＴ９以降に発生する。時刻Ｔ１０において、パリティ・ブロックＰ０への書込み命令がドライブ３のコマンド・キューの入力端に配置される。時刻Ｔ１０の時点で、パリティ・ブロックＰ０に対する２つの書込み命令がドライブ３のコマンド・キュー内に同時に存在する。時間間隔Ｔ１１及びＴ１２では、パリティ・ブロックＰ０に対する２番目の書込み命令がドライブ３のコマンド・キュー内において進行する。

【００２９】図５はパリティ・キャッシュ・テーブル３１の構造を表す。パリティ・キャッシュ・テーブルは、メモリ・バッファ２７に存在するパリティ・データ・ブロックに関連する最大Ｋ個のエントリを含む。別の実施例では、パリティ・キャッシュ・テーブル３１及びパリティ・データ・ブロック３３が、システム・メモリ１５に記憶される。当業者には、メモリ・バッファ２７内のデータへのアクセスが、システム・メモリ１５へのアクセスよりも高速であることが望ましいことが理解されよう。いずれにしても、これらのいずれのアクセスもディスク・ドライブからの回復よりは高速である。データ構造を記憶するためのシステム・メモリ１５の使用により、通信リンク１９上のトラフィックは増加する。

【００３０】パリティ・キャッシュ・テーブル３１内の各エントリは、ＲＡＩＤレベル４システムのドライブ・ユニット４１Ｄ、またはＲＡＩＤレベル５システムのドライブ４１Ａ乃至４１Ｄの１つにおける論理ブロック・アドレス範囲に対応する論理ブロック・アドレス範囲を含む。アドレス範囲フィールド７０は、ドライブ・ユニット上の開始論理ブロック・アドレス及びパリティ・ブロックの終りを含む。エントリには更に、エントリに関連するパリティ・データへの最も最近の使用またはアクセスのタイム・スタンプ７１が含まれる。NWW（number of writeswaiting）フィールド７３は、実行を待機中の書込みの数を示すカウンタである。フィールド７５（Ｌ／Ｆ）はデータへのアクセスをロックするためのフラグであり、メモリ・バッファ２７内のパリティ・データの直列化の変更を提供する。最後に、データ・キャッシュ・テーブル３１内のエントリは、パリティ・データ・ブロックを見い出すためのバッファ内における位置を示す。データ・キャッシュ・テーブル３１及びパリティ・データ・ブロック３３は、電力損失を防ぐために、不揮発メモリ内に保持されてもよい。

【００３１】図６は、ホスト処理ユニットからの更新コマンド、すなわちＲＡＩＤ２１へのデータの読出し－変更－書込みを要求するコマンドの受信において入力される処理の論理流れ図である。付加データの識別を含む更新コマンドの受信に際し、適切なストライプのデータ及びパリティ・データに対応する論理ブロック・アドレス範囲がステップ２０１で計算される。プログラムはユーザ・データの更新及びパリティ・データの更新のための処理を両立するが、これは実質的には非同期処理であり、ドライブが非同期に動作する。ユーザ・データ処理ではステップ２０３へ移行し、適切なデータ・ドライブ・コマンド・キューがフルかどうかを判断する。フルの場合、処理はエントリがキュー内で開くまで待機する。ユーザ・データ・ドライブのコマンド・キューの空間が使用可能になると、分岐せずにステップ２０３からステップ２０５に移行する。ステップ２０５で、１）ステップ２０１で決定された論理ブロック・アドレスから旧データを読出し、２）付加データを同一位置に書込むために、リンク・コマンドがデータ・ドライブ・コマンド・キューに送信される。処理は次にステップ２０７に移行し、付加データと置換パリティ・データの両方が書込まれたことが示されるまで、待機オペレーションを実行する。ステップ２０７は実質的に２つの処理の再同期化を提供する。ステップ２０７の後、ステップ２０９が実行され、完全なステータスをホストに送信する。処理はこの時点で終了する。

【００３２】パリティ・データ処理ではステップ２１１が実行され、パリティ・ドライブ・キューがフルかどうかが判断される。フルの場合、パリティ・データを保持するドライブのコマンド・キューに、コマンドのための空間が使用可能になるまで、待機が実行される。空間が使用可能となると分岐せずにステップ２１３に移行す

る。ステップ２１３では、ステップ２０１でパリティ・データに対して計算されたパリティ・キャッシュ・テーブルの論理ブロック・アドレス範囲に、エントリが存在するかどうかが判断される。パリティ・キャッシュ・ヒットが発生すると、ステップ２１３からステップ２１５へ分岐する。ステップ２１５では、パリティ・データ・ブロックのステータスがフリーかどうかが判断される。フリーでない場合、ステータス・フラグがフリーになるまで待機が実行される。次にステップ２１５からステップ２１７に分岐し、ステータス・フラグをロックにセットし、待機中の書込み数を増分し、パリティ・キャッシュ・エントリ内のタイム・スタンプを置換する。ステップ２１７に続きステップ２１９が実行され、新たなまたは置換パリティを計算するために要求される旧データが、ステップ２０５から使用可能かどうかを判断する。処理はステップ２１９で旧ユーザ・データが使用可能となるまで待機する。データが使用可能になると、ステップ２２１に分岐し、新たなパリティ・データを計算し、キャッシュ内のパリティ・ブロックを更新する。ステップ２２３でステータス・フラグがフリーにセットされ、ステップ２２５で新たなパリティ・データの書込みコマンドが適切なディスク・ドライブに送信され、その入出力制御装置に対応するコマンド・キューに待機される。ステップ２２７は、入出力制御装置がパリティ書込みオペレーションの完了を示すまで待機サイクルを提供する。書込みが完了するとステップ２２９が実行され、待機中の書込み数が１減分される。処理は次にステップ２０７に移行する。

【００３３】ステップ２１３でパリティ・キャッシュ・ヒットが発生しなかったと判断されると、分岐せずにステップ２３１に移行する。ステップ２３１では、フリー・パリティ・キャッシュ・エントリが使用可能かどうかが判断される。使用可能な場合、ステップ２３３で、フリー・パリティ・キャッシュ・エントリを使用し、新たなパリティ・キャッシュ・エントリが作成される。使用可能でない場合には、ステップ２３５でパリティ・キャッシュ・エントリがテーブルから解放される。これは待機中の書込み数が０である最低使用頻度（ＬＲＵ）パリティ・キャッシュ・エントリを見い出すことにより実行される。エントリが探し出されると、ステップ２３３で新たなパリティ・エントリが、解放されたエントリから作成される。次にステップ２３７でステータス・フラグがロックにセットされ、待機中の書込み数が１にセットされる。次にステップ２３９で旧パリティがパリティ・ドライブから読出される。これはステップ２４１で読出しオペレーションの完了として示されるように、コマンドの通常のキューイング及び待機を含むものと理解される。パリティの読出しが完了し、その読出されたばかりのパリティのバッファ２７における記憶位置がキャッシュ・エントリに含まれると、ステップ２４１からステップ２１９に分岐し、処理が上述のように継続される。

【００３４】入出力制御装置レベルにおけるタグ・コマンド・キューイングによる性能の改良は、こうしたシステムが本発明を使用するＲＡＩＤレベル４またはレベル５大容量記憶システムに適用されると失われることはない。キャッシュ・エントリ・テーブルのデータ構造が、電源障害による損失から保護するために不揮発ＲＡＭにより実現されてもよい。

【００３５】本発明は特定の実施例に関連して述べられてきたが、当業者には本発明の精神及び範囲から逸脱することなく、形態及び詳細に関する様々な変更が可能であることが理解されよう。

【００３６】まとめとして、本発明の構成に関して以下の事項を開示する。

【００３７】（１）耐障害論理大容量記憶装置内の物理記憶装置にユーザ・データ及び冗長データを記憶する方法であって、ホストからの付加データを伴う書込みコマンドの受信に応答して、第１の物理記憶装置上における付加データの論理ブロック・アドレスと、第２の物理記憶装置上における冗長データのパリティ・アドレスとを決定するステップと、前記第１の記憶装置に対して、論理ブロック・アドレスの旧データを読出し、論理ブロック・アドレスに付加データを書込むように指令するステップと、前記第２の記憶装置上の冗長データがメモリ・バッファ内に存在するかどうかを判断するステップと、冗長データが前記メモリ・バッファ内に存在しない場合、前記第２の記憶装置から前記メモリ・バッファに冗長データを読出すステップと、冗長データ、付加データ及び旧データから新たな冗長データを計算するステップと、前記第２の物理記憶装置に新たな冗長データを論理ブロック・アドレスに記憶するように指令するステップと、を含む記憶方法。

（２）新たな冗長データを計算するステップが、前記メモリ・バッファ内の冗長データを新たな冗長データにより置換するステップを含む、前記（１）記載の記憶方法。

（３）ディスク・ドライブの冗長アレイ内のパリティ・ドライブ上にコマンドをキューイングする方法であって、ホストからの付加データを伴う書込みコマンドの受信に応答して、データ・ドライブ上における付加データの論理ブロック・アドレスと、前記パリティ・ドライブ上におけるパリティ・データの論理ブロック・アドレスとを決定するステップと、パリティ・キャッシュ・エントリがパリティ・データの論理ブロック・アドレスに一致するかどうかを判断するステップと、前記一致の肯定判定に応答して、パリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションのパリティ・データと、前記データ・ドライブから読出されるデータと付加データとから、置換パリティ・データを計算するステップと、置換パリティ・データを前記バッファ・ロケー

ションに配置するステップと、置換パリティ・データを書込むためのコマンドを、パリティ・データの論理ブロック・アドレスに対応する入出力制御装置に送信するステップと、を含むキューイング方法。

（4）前記一致の否定判定に応答して、パリティ・データを前記バッファ・ロケーションに読出すためのコマンドを、前記パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信するステップと、前記パリティ・ドライブの論理ブロック・アドレスのパリティ・キャッシュ・エントリを生成するステップと、を含む、前記（3）記載のキューイング方法。

（5）パリティ・キャッシュ・エントリが前記パリティ・ドライブの論理ブロック・アドレスと、ドライブ識別と、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、メモリ・バッファ内のパリティ・データの位置を示す位置フィールドとを含む、前記（4）記載のキューイング方法。

（6）前記データ・ドライブの論理ブロック・アドレスの決定に応答して、データを読出し次に付加データを書込むためのリンク・コマンドを、論理ブロック・アドレスをカバーする入出力制御装置に送信するステップ、を含む、前記（5）記載のキューイング方法。

（7）計算ステップがリンク・コマンドの送信ステップの読出しオペレーションの後に発生する、前記（6）記載のキューイング方法。

（8）前記パリティ・キャッシュ・エントリの生成ステップが、パリティ・キャッシュ・テーブルが未使用のパリティ・キャッシュ・エントリを有するかどうかを判断するステップと、肯定判定に応答して、未使用のパリティ・キャッシュ・エントリを使用するステップと、否定判定に応答して、待機中の書込みを有さないパリティ・キャッシュ・エントリの中から、最低使用頻度のパリティ・キャッシュ・エントリを探し出し、探し出されたパリティ・キャッシュ・エントリを使用するステップと、を含む、前記（5）記載のキューイング方法。

（9）置換パリティ・データ書込みのためのコマンドの送信ステップが、書込み待機カウンタを増分するステップを含む、前記（5）記載のキューイング方法。

（10）置換パリティ・データの生成のために、前記バッファ・ロケーション内のパリティ・データへのアクセスを直列化するステップ、を含む、前記（9）記載のキューイング方法。

（11）置換パリティ・データが前記パリティ・ドライブの論理ブロック・アドレスに書込まれた後、パリティ・キャッシュ内の書込み待機カウンタを減分するステップ、を含む、前記（10）記載のキューイング方法。

（12）前記パリティ・ドライブがチェック・ディスクである、前記（11）記載のキューイング方法。

（13）前記パリティ・ドライブがディスク・ドライブ

の冗長アレイの任意の1つに相当する、前記（11）記載のキューイング方法。

（14）ホスト・データ処理システムからの付加データを伴う書込み命令の受信に応答して、データ・ドライブ上における付加データの論理ブロック・アドレスと、パリティ・ドライブ上における冗長データの論理ブロック・アドレスとを決定する手段と、複数のパリティ・キャッシュ・エントリを有するパリティ・キャッシュ・テーブルと、パリティ・キャッシュ・エントリ及び前記パリティ・ドライブ上の論理ブロック・アドレスに対応するパリティ・データを記憶するバッファと、パリティ・キャッシュ・エントリが発見手段により見い出される前記パリティ・ドライブの論理ブロック・アドレスに一致するかどうかを判断する手段と、前記一致の肯定判定に応答して、パリティ・キャッシュ・エントリにより提供されるバッファ・ロケーションのパリティ・データと、前記データ・ドライブの論理ブロック・アドレスから読出されるデータと、付加データとから、置換パリティ・データを計算する手段と、置換パリティ・データを前記バッファ・ロケーションに配置する手段と、置換パリティ・データを書込むためのコマンドを、パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段と、を含む、ディスク・ドライブの冗長アレイ。

（15）前記一致の否定判定に応答して、パリティ・データを前記バッファ・ロケーションに読出すためのコマンドを前記パリティ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段と、前記パリティ・ドライブの論理ブロック・アドレスから読出されるパリティ・データに対応するパリティ・キャッシュ・エントリを生成する手段と、を含む、前記（14）記載のディスク・ドライブの冗長アレイ。

（16）パリティ・キャッシュ・エントリが前記パリティ・データの論理ブロック・アドレスと、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、ドライブ識別と、メモリ・バッファ内のパリティ・データの位置フィールドとを含む、前記（15）記載のディスク・ドライブの冗長アレイ。

（17）前記データ・ドライブの論理ブロック・アドレスの決定に応答して、データを読出し次に付加データを書込むためのリンク・コマンドを、前記データ・ドライブの論理ブロック・アドレスに対応する入出力制御装置に送信する手段、を含む、前記（16）記載のディスク・ドライブの冗長アレイ。

（18）パリティ・キャッシュ・エントリの生成手段が、パリティ・キャッシュ・テーブルが未使用のパリティ・キャッシュ・エントリを有するかどうかを判断する手段と、肯定判定に応答して、未使用のパリティ・キャッシュ・エントリを使用する手段と、否定判定に応答し

て、待機中の書込みを有さない最低使用頻度のパリティ・キャッシュ・エントリを探し出し、探し出されたパリティ・キャッシュ・エントリを使用する手段と、を含む、前記（１６）記載のディスク・ドライブの冗長アレイ。

（１９）置換パリティ・データ書込みのためのコマンドの送信手段が、書込み待機カウンタを増分する手段を含む、前記（１６）記載のディスク・ドライブの冗長アレイ。

（２０）置換パリティ・データの生成のために、前記バッファ・ロケーション内のパリティ・データへのアクセスを直列化する手段、を含む、前記（１９）記載のディスク・ドライブの冗長アレイ。

（２１）置換パリティ・データが前記パリティ・ドライブの論理ブロック・アドレスに書込まれた後、前記パリティ・ドライブの論理ブロック・アドレスに対応するパリティ・キャッシュ内の書込み待機カウンタを減分する手段、を含む、前記（２０）記載のディスク・ドライブの冗長アレイ。

（２２）前記パリティ・ドライブがチェック・ディスクである、前記（２１）記載のディスク・ドライブの冗長アレイ。

（２３）前記パリティ・ドライブがディスク・ドライブの冗長アレイの任意の１つに相当する、前記（２１）記載のディスク・ドライブの冗長アレイ。

（２４）論理装置として動作する物理記憶装置のアレイと、物理記憶装置のアレイ上の所定の位置に記憶される付加データを伴う更新コマンドを提供するコマンド発生器と、ユーザ・データのストライプに対応する選択パリティ・データ・ブロックのコピーを記憶するキャッシュと、ユーザ・データ・ブロック及びユーザ・データ・ブロックのストライプに渡って生成されるパリティ・データ・ブロックを物理記憶装置のアレイ間でストライプし、物理記憶装置位置によりキャッシュ内に記憶される選択パリティ・データ・ブロックの１個以上のコピーを識別するパリティ・キャッシュ・テーブルを保守するローカル・プロセッサと、ローカル・プロセッサによりそれぞれの入出力制御装置に送信されるコマンドをキューイングする、物理記憶装置の物理アレイの各々に関連する入出力制御装置と、を含むデータ処理システム。

（２５）パリティ・キャッシュ・エントリが物理記憶装置の論理ブロック・アドレスと、物理記憶装置の識別と、タイム・スタンプと、ロック／非ロック・フラグと、待機中の書込みの数を示す書込み待機カウンタと、キャッシュ内のパリティ・データ・ブロックの位置・フィールドとを含む、前記（２４）記載のデータ処理システム。

（２６）ローカル・プロセッサが、キャッシュ内の選択パリティ・データ・ブロックの各コピーへのアクセスを制御するシリアライザ、を含む、前記（２５）記載のデータ処理システム。

（２７）キャッシュ内にパリティ・データ・ブロックが存在しないストライプ内の物理アドレスを更新するコマンドに応答して、前記ストライプに対応するパリティ・データ・ブロックをキャッシュに読出し、パリティ・キャッシュ・テーブルに前記パリティ・データ・ブロックのエントリを生成するパリティ・キャッシュ・テーブル・エントリ発生器、を含む、前記（２６）記載のデータ処理システム。

（２８）パリティ・データが単一の物理記憶装置上に配置される、前記（２７）記載のデータ処理システム。

（２９）パリティ・データが全ての物理記憶装置上に配置される、前記（２７）記載のデータ処理システム。

【００３８】

【発明の効果】以上説明したように、本発明によれば、集合的に１個以上の論理大容量記憶装置として機能する大容量記憶装置のセットが提供され、ＲＡＩＤレベル４及び５システムにおけるパリティ・ドライブ上で、コマンドをキューイングするシステム及び方法が提供される。

【図面の簡単な説明】

【図１】ＲＡＩＤを含むデータ処理システムのハイレベル・ブロック図である。

【図２】ディスク・ドライブの上面図である。

【図３】ＲＡＩＤレベル４及び５システムにおけるデータ・ストライピングを表す図である。

【図４】ＲＡＩＤ４または５システムにおいてストライプされるデータを変更するコマンド・シーケンスのタイミング図である。

【図５】本発明をサポートするために使用されるパリティ・キャッシュ・テーブルのデータ構造を表す図である。

【図６】本発明を実施するために、ＲＡＩＤシステムの記憶制御装置上で実行される処理のハイレベル論理流れ図である。

【符号の説明】

１１　データ処理システム
１３　中央処理ユニット（ＣＰＵ）
１５　システム・メモリ
１７　大容量記憶制御装置
１９　通信リンク
２１　ＲＡＩＤ
２３　インタフェース
２５　ローカル・プロセッサ
２７　メモリ・バッファ
２９　排他的論理和プロセッサ
３１　パリティ・キャッシュ・テーブル
３３　パリティ・データ・ブロック
３５　ユーザ・データ・ブロック
４５　オペレーティング・システム

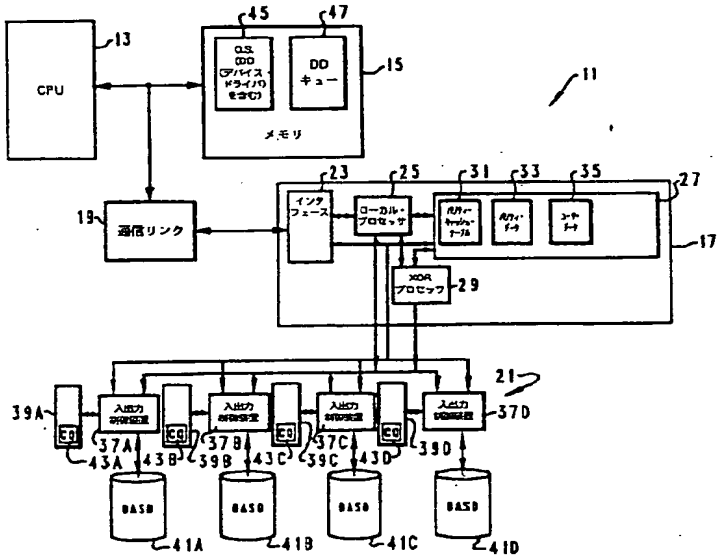
21

４７　デバイス・ドライバ・キュー
７０　アドレス範囲フィールド
７１　タイム・スタンプ
７３　NWWフィールド
７５　フィールド
１２０　ディスク・ドライバ

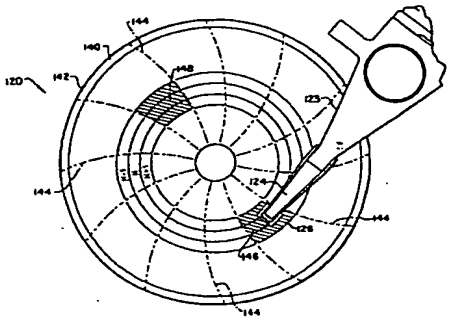
22

１２３　アーム
１２６　スライダ
１４０　磁気表面
１４２　ディスク
１４４　トラック・サーボ・フィールド
１４６、１４８　セクタ

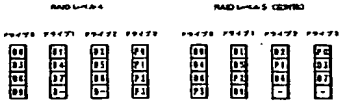
【図１】



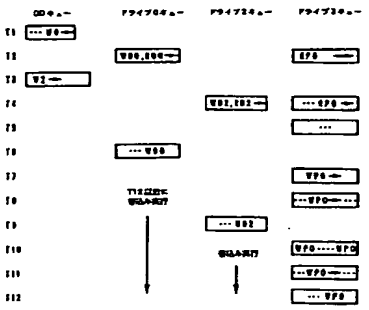
【図２】



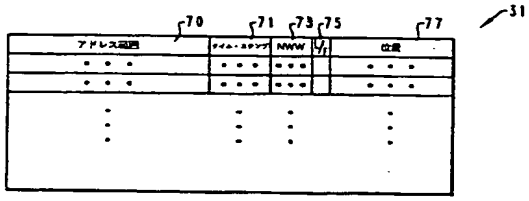
【図３】



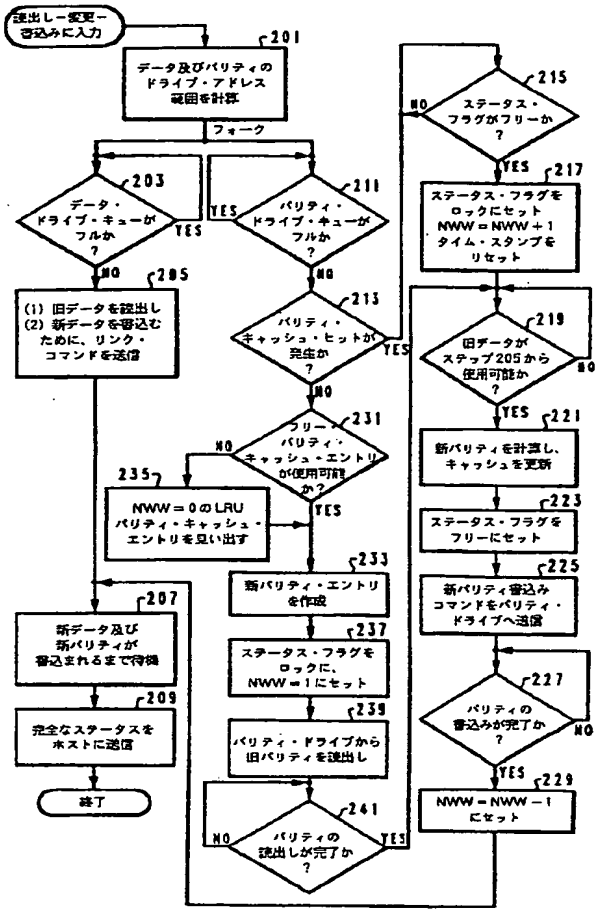
【図4】



【図5】



【図6】


読出し－変更－書込みに入力
201
データ及びパリティのドライブ・アドレス範囲を計算
フォーク
203
データ・ドライブ・キューがフルか？
YES
NO
205
(1) 旧データを読出し (2) 新データを書込むために、リンク・コマンドを送信
211
パリティ・ドライブ・キューがフルか？
YES
NO
213
パリティ・キャッシュ・ヒットが発生か？
YES
NO
231
フリー・パリティ・キャッシュ・エントリが使用可能か？
NO
YES
235
NWW＝0のLRUパリティ・キャッシュ・エントリを見い出す
233
新パリティ・エントリを作成
237
ステータス・フラグをロックに、NWW＝1にセット
239
パリティ・ドライブから旧パリティを読出し
241
パリティの読出しが完了か？
NO
YES
207
新データ及び新パリティが書込まれるまで待機
209
完全なステータスをホストに送信
終了
215
ステータス・フラグがフリーか？
NO
YES
217
ステータス・フラグをロックにセット NWW＝NWW＋1 タイム・スタンプをリセット
219
旧データがステップ205から使用可能か？
NO
YES
221
新パリティを計算し、キャッシュを更新
223
ステータス・フラグをフリーにセット
225
新パリティ書込みコマンドをパリティ・ドライブへ送信
227
パリティの書込みが完了か？
NO
YES
229
NWW＝NWW－1にセット